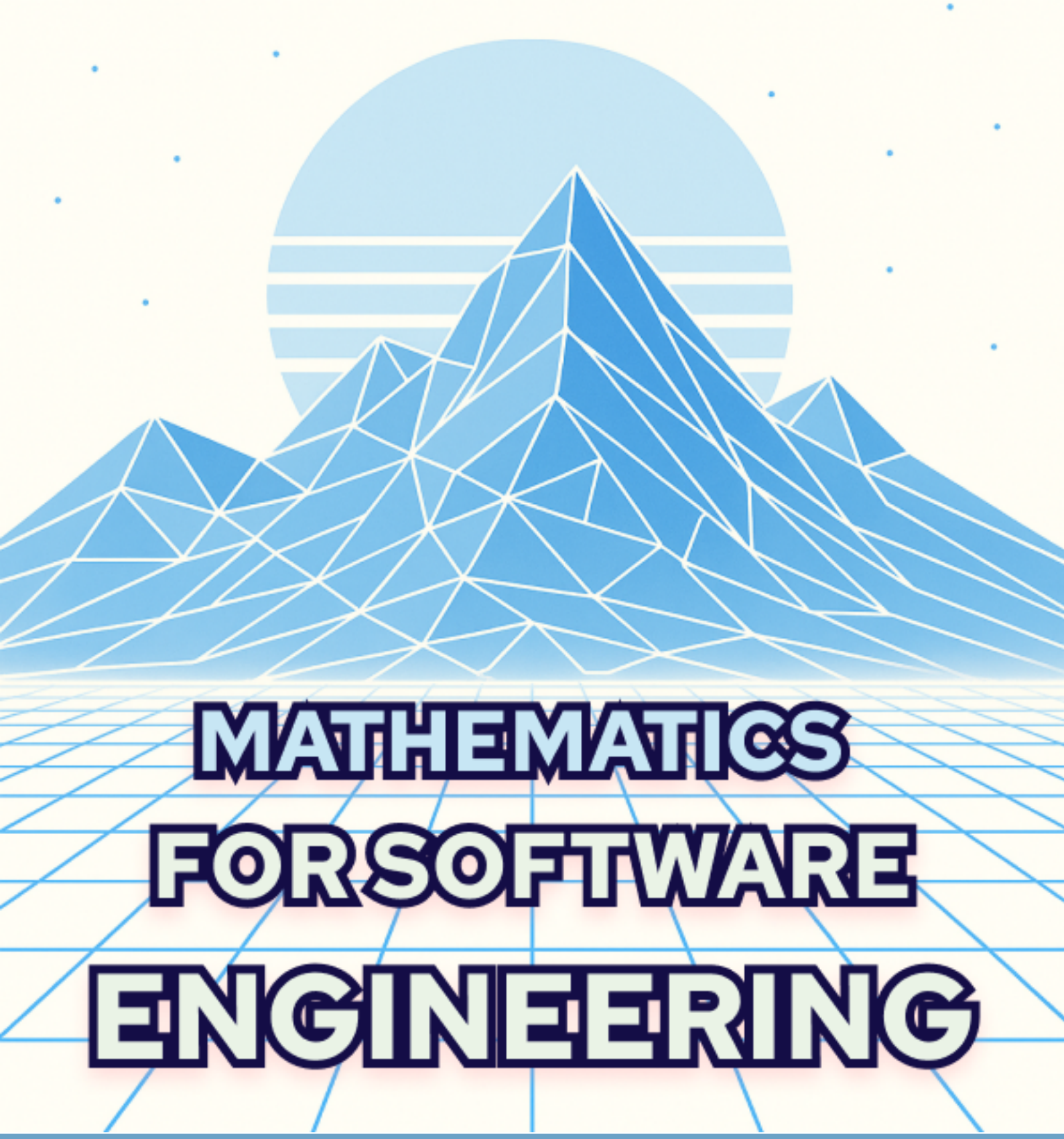


Authors: Richard Brooks & Eduard Fekete

Institute: VIA University College

Date: May, 2025

Version: 2.0

Contents

Part I

Differential Calculus

Chapter 1 Differential Calculus

In both practical experience and modern engineering, we frequently encounter situations where one quantity changes in response to another. A car's fuel consumption reflects how distance depends on the amount of fuel used. A sprinter's velocity describes how position varies with time. In agriculture, yield changes with the amount of fertiliser, and in economics, demand responds to price. The study of such relationships lies at the heart of *differential calculus*, the mathematical discipline devoted to understanding how quantities vary together.

The concept that captures this variation is the *rate of change*.

- The change of distance with respect to time is *velocity*.
- The change of profit with respect to price is *marginal profit*.
- The change of loss with respect to model parameters is the *gradient*.

When a rate of change remains constant, the relationship between the quantities is linear, and the graph forms a straight line. When the rate of change varies, the graph becomes curved, and the slope depends on the specific point of measurement. This naturally leads to the idea of a *tangent line*, which locally approximates a curve at a single point.

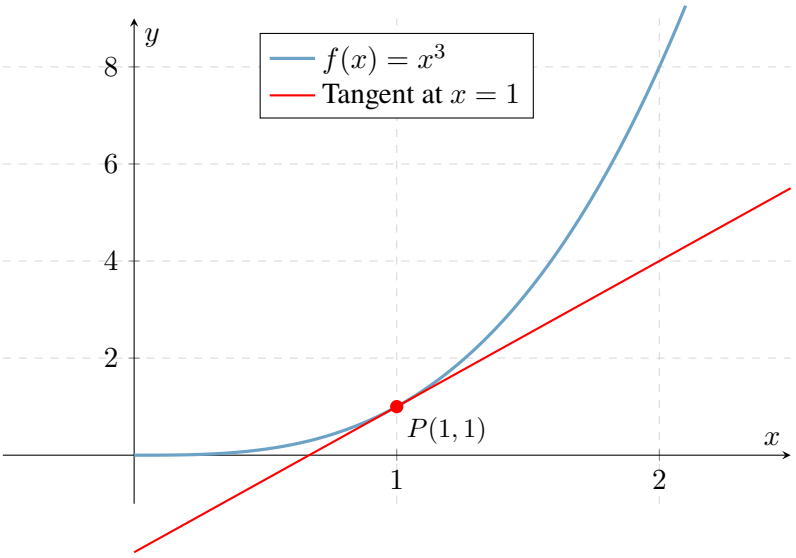


Figure 1.1: The tangent line to the graph of $f(x) = x^3$ at the point $P(1, 1)$.

Remark: For a straight line, the slope is constant. For a curved graph, the slope varies from point to point, so we speak of the *instantaneous rate of change*, the slope of the tangent at a specific point.

The same principle extends to computational and data-driven systems. A program's runtime changes with input size, a model's accuracy varies with the number of training epochs, and network throughput depends on available bandwidth. In machine learning, the process of training a model resembles a hiker searching for the lowest point in a vast, foggy landscape. The "altitude" represents the model's error, and progress requires moving downhill. Determining the correct direction among millions of possible ones depends on knowing the local slope — the *derivative* — which expresses the instantaneous rate of change.

Differential calculus provides the tools to describe, compute, and interpret such changes with precision. It

transforms intuitive ideas such as being faster, steeper, or more efficient into exact mathematical expressions. This chapter introduces the fundamental ideas of differential calculus, beginning with the concepts of *limits* and *continuity*, which formalise the notion of approaching a point infinitely closely. Building on these, we develop the definition and interpretation of the *derivative* as a measure of instantaneous change and explore its applications in contexts ranging from physical motion to the optimisation of modern algorithms.

1.1 Limits and Continuity: The Foundation of Calculus

Before we can measure an instantaneous rate of change, we must first build a formal understanding of what it means to get "infinitesimally close" to a point without necessarily being at that point. This is the concept of a **limit**. Building on limits, **continuity** gives us the language to describe functions that are predictable and well-behaved, without sudden jumps or breaks.

Limits

A limit describes the value that a function "approaches" as its input gets closer and closer to a specific point. This idea is fundamental not just in calculus, but in software engineering, where we often analyze the behavior of systems under approaching conditions.

For example, when analyzing an algorithm's runtime, we ask what happens to the execution time as the input size 'n' approaches infinity. In networking, we might model the theoretical maximum throughput as a limit when latency approaches zero.

Definition 1.1 (Limit)

Let $f(x)$ be a function defined near a point c. We say that the **limit** of $f(x)$ as x approaches c is L, written as

$$\lim_{x \to c} f(x) = L$$

if we can make the value of $f(x)$ arbitrarily close to L by choosing an x that is sufficiently close to c, but not equal to c.

In many simple cases, the limit is just the value of the function at that point. For example, for $f(x) = x^2$, the limit as x approaches 3 is simply $3^2 = 9$. The real power of limits, however, is in handling cases where the function is undefined at the point of interest.

Example 1.1 A Limit at an Undefined Point

Consider the function $f(x) = \frac{\sin(x)}{x}$. This function is fundamental in signal processing and is known as the sinc function. Notice that $f(0)$ is undefined because it results in division by zero. However, we can ask what value the function approaches as x gets very close to 0.

By plugging in values very close to 0, we can observe a trend:

$$f(0.1) = \frac{\sin(0.1)}{0.1} \approx 0.9983$$
$$f(0.01) = \frac{\sin(0.01)}{0.01} \approx 0.99998$$
$$f(-0.01) = \frac{\sin(-0.01)}{-0.01} \approx 0.99998$$

The function appears to approach 1. In calculus, it can be formally proven that:

$$\lim_{x \to 0} \frac{\sin(x)}{x} = 1$$

This ability to analyze behavior at a problematic point is essential for creating robust numerical algorithms.

Example 1.2 Asymptotic Analysis with Limits

Suppose you want to analyze whether the function $f(n) = 5n^2 + 4n + 7$ is $\mathcal{O}(n^2)$ as n grows large, as is common in the study of algorithm complexity. One way to formalize this is to examine the limit:

$$\lim_{n \to \infty} \frac{f(n)}{n^2} = \lim_{n \to \infty} \frac{5n^2 + 4n + 7}{n^2} = \lim_{n \to \infty} \left(5 + \frac{4}{n} + \frac{7}{n^2}\right)$$

As n becomes very large, the terms $\frac{4}{n}$ and $\frac{7}{n^2}$ both approach zero, so the whole expression approaches 5. Because this limit is a finite constant, it confirms that $f(n)$ grows at the same rate as n^2. This type of limit calculation is at the heart of asymptotic analysis, formally justifying Big-$\mathcal{O}$ claims about algorithm runtime.

Continuity

Continuity formalizes the idea of a function being "well-behaved" or predictable. Intuitively, a function is continuous if you can draw its graph without lifting your pen from the paper. In software, we often assume our systems are continuous: a small change in input should lead to a small, predictable change in output, not a sudden, drastic jump.

Definition 1.2 (Continuity)

A function f is **continuous** at a point c if the following three conditions are met:

1. $f(c)$ is defined.
2. $\lim_{x \to c} f(x)$ exists.
3. $\lim_{x \to c} f(x) = f(c)$.

A function is continuous on an interval if it is continuous at every point in that interval. ♣

Discontinuities often represent important events in software systems, such as state transitions or threshold triggers.

Example 1.3 Jump Discontinuity in a SaaS Pricing Model

Consider a cloud service that charges based on the number of API calls per month. The pricing model is tiered: \$20 for up to 10,000 calls, and \$40 for any usage above 10,000 calls. The cost function $C(x)$, where x is the

number of calls, has a **jump discontinuity** at $x = 10,000$.

$$C(x) = \begin{cases} 20 & \text{if } x \leq 10000 \\ 40 & \text{if } x > 10000 \end{cases}$$

As shown in Figure 1.2, the cost suddenly jumps at the threshold. Understanding discontinuities is crucial for modeling any system with threshold-based logic, from auto-scaling policies to billing systems.

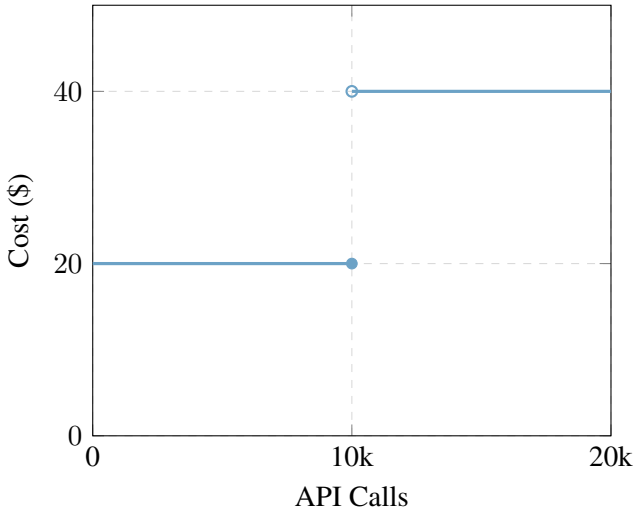


Figure 1.2: The cost function exhibits a jump discontinuity at 10,000 API calls as the price tier changes.

1.2 The Derivative: Measuring the Rate of Change

We often talk about average rates of change. For example, if a data transfer takes 10 seconds and moves 500 MB, the average transfer rate is 50 MB/s. But this average tells us nothing about fluctuations during the transfer. The **derivative** is the tool that lets us move from this average rate to the **instantaneous rate of change** at any specific moment.

We may thus say that differential calculus is about determining how one quantity changes in relation to another. Geometrically, the derivative of a function at a point is the slope of the line tangent to the function's graph at that point. To find the slope of a tangent, we approximate it using secant lines between two nearby points on the curve. As the points move closer together, the secant slope approaches the tangent slope. As shown in Figure 1.3, this tangent line is the limit of secant lines passing through two points on the curve as the points get infinitesimally close.

In Figure 1.3 as the point Q moves along the curve toward P (from Q_1 to Q_2 to P), the slope of the secant line PQ approaches the slope of the tangent line to $f(x) = x^3$ at P. Note how the tangent only touches the curve at P, while the secant passes through two points.

This leads to the formal definition of the derivative.

Definition 1.3 (The Derivative)

The **derivative** of a function f with respect to x, denoted $f'(x)$, is the function

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

provided the limit exists. If $f'(x)$ exists, we say that f is **differentiable** at x.

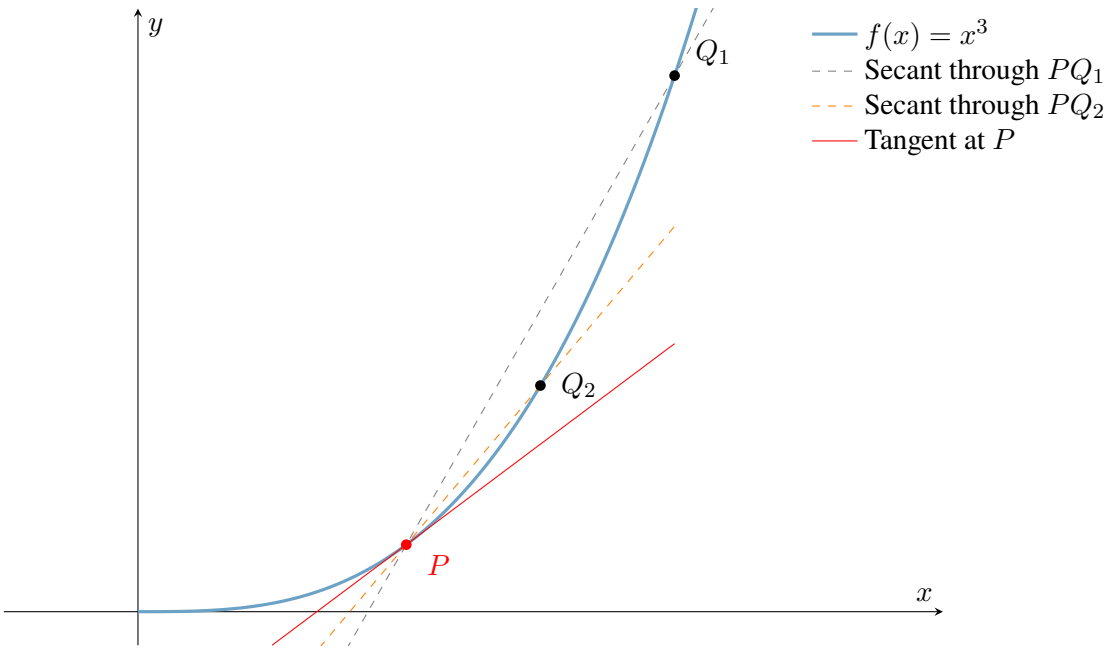


Figure 1.3: Illustration of the tangent line as a limit of secant lines.

The expression $\frac{f(x+h)-f(x)}{h}$ is the **difference quotient**, representing the average rate of change over the interval from x to $x+h$. The derivative is the limit of this average rate as the interval size h shrinks to zero.

Example 1.4 Derivative from First Principles

Let's find the derivative of $f(x) = x^2$ using the limit definition.

Solution: We start with the difference quotient and simplify:

$$\frac{f(x+h)-f(x)}{h} = \frac{(x+h)^2 - x^2}{h} = \frac{x^2 + 2xh + h^2 - x^2}{h} = \frac{2xh + h^2}{h} = 2x + h$$

Finally, we take the limit as $h \to 0$:

$$f'(x) = \lim_{h \to 0} (2x + h) = 2x$$

Thus, the derivative of $f(x) = x^2$ is $f'(x) = 2x$. This tells us the slope of the parabola at any point x. ◀

Notation

The derivative of f can be written in several equivalent forms:

$$f'(x), \quad \frac{df}{dx}, \quad \frac{dy}{dx}, \quad Df(x), \quad \frac{d}{dx}f(x).$$

1.3 Rules of Differentiation

Differential calculus provides a way to find the exact derivative of a function directly from its formula, without relying on graphs or numerical estimation. In practice, this is done using a set of simple rules that allow us to compute the derivative of nearly any function we are likely to encounter. In this section, we will introduce these rules, explain their meaning, and show how to apply them in practice.

Constant and Power Rules

Perhaps the simplest functions in mathematics are the constant functions and the functions of the form x^n.

Theorem 1.1 (Constant and Power Rules)

1. **Constant Rule:** If c is constant, then $\dfrac{d}{dx}(c) = 0$,
2. **Power Rule:** For any real n, $\dfrac{d}{dx}x^n = nx^{n-1}$.

♡

Example 1.5

- If $f(x) = x^7$, then $f'(x) = 7x^6$,
- If $y = x^{-0.5}$, then $\frac{dy}{dx} = -0.5x^{-1.5}$,
- $\frac{d}{dx}x^{-3} = -3x^{-4}$,
- If $g(x) = 3.2$, then $g'(x) = 0$,
- If $f(t) = \sqrt{t} = t^{\frac{1}{2}}$, then $f'(t) = \frac{1}{2}t^{-\frac{1}{2}} = \frac{1}{2\sqrt{t}}$,
- If $h(u) = -13.29$, then $h'(u) = 0$,
- $\dfrac{d}{dx}x^3 = 3x^2$, $\dfrac{d}{dx}x^{-2} = -2x^{-3}$.

In the examples above, we have used the constant and power rules to find the derivatives of several simple functions. However, it is important to remember what these calculations actually represent. The derivative tells us the slope of the tangent line to the graph of a function at any given point.

For instance, if $f(x) = x^2$, then $f'(x) = 2x$. To find the slope of the tangent to the graph of x^2 at $x = 1$, we substitute $x = 1$ into the derivative, giving $f'(1) = 2 \times 1 = 2$. Similarly, the slope of the tangent at $x = -0.5$ is $f'(-0.5) = 2 \times (-0.5) = -1$.

These results are illustrated in Figure 1.4, where the tangent lines at $x = 1$ and $x = -0.5$ have slopes of 2 and -1, respectively.

Example 1.6

Let $f(x) = (x^2 + 1)^3$. Then

$$f'(x) = 3(x^2 + 1)^2 \cdot 2x = 6x(x^2 + 1)^2.$$

Example 1.7

Find the slope of the tangent to the graph of the function $g(t) = t^4$ at the point on the graph where $t = -2$.

Solution:

The derivative is $g'(t) = 4t^3$, and so the slope of the tangent line at $t = -2$ is $g'(-2) = 4 \times (-2)^3 = -32$. ◀

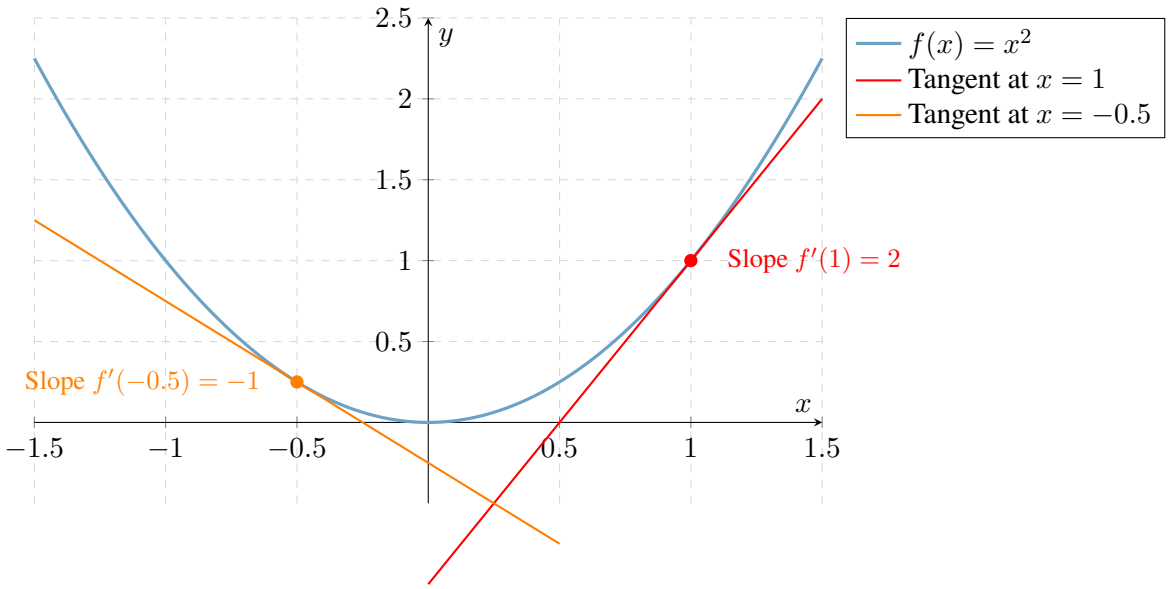


Figure 1.4: The tangent lines to the graph of $f(x) = x^2$ at different points have different slopes.

Example 1.8

Find the equation of the line tangent to the graph of $y = f(x) = x^{\frac{1}{2}}$ at the point $x = 4$.

Solution:

$f(4) = 4^{\frac{1}{2}} = \sqrt{4} = 2$, so the coordinates of the point on the graph are $(4, 2)$. The derivative is

Any non vertical line has equation of the form $y = mx + b$ where m is the slope and b the vertical intercept.

In this case the slope is $\frac{1}{4}$, so $m = \frac{1}{4}$, and the equation is $y = \frac{x}{4} + b$. Because the line passes through the point $(4, 2)$ we know that $y = 2$ when $x = 4$.

Substituting we get $2 = \frac{4}{4} + b$, so that $b = 1$. The equation is therefore $y = \frac{x}{4} + 1$. ◀

We now know how to differentiate any function that is a power of the variable. Examples are functions like x^3 and $t^{-1.3}$. You will come across functions that do not at first appear to be a power of the variable, but can be rewritten in this form. One of the simplest examples is the function

$$f(t) = \sqrt{t}$$

which can also be written in the form

$$f(t) = t^{\frac{1}{2}}.$$

The derivative is then

$$f'(t) = \frac{t^{-\frac{1}{2}}}{2} = \frac{1}{2\sqrt{t}}$$

Similarly, if

$$h(s) = \frac{1}{s} = s^{-1}$$

then

$$h'(s) = -s^{-2} = -\frac{1}{s^2}$$

Example 1.9

If $f(x) = \frac{1}{\sqrt[3]{x}} = x^{-\frac{1}{3}}$ then $f'(x) = -\frac{1}{3}x^{-\frac{4}{3}}$.

Example 1.10

If $y = \frac{1}{x\sqrt{x}} = x^{-\frac{3}{2}}$ then $\frac{dy}{dx} = -\frac{3}{2}x^{-\frac{5}{2}}$.

Scalar Rule and Linearity

So far we know how to differentiate powers of the independent variable. Many of the functions that arise in applications are built from such powers in simple ways. For instance, the function $3x^2$ is merely a scalar multiple of x^2, yet neither the constant and power rules tell us how to differentiate $3x^2$. Likewise, these rules do not explain how to differentiate expressions such as $x^2 + x^3$ or $x^2 - x^3$.

Theorem 1.2 (Scalar and Linearity Rules)

3. **Scalar Rule:** $\frac{d}{dx}[c \cdot f(x)] = c \cdot f'(x)$.
4. **Linearity Rule:** $\frac{d}{dx}[f(x) \pm g(x)] = f'(x) \pm g'(x)$.

♡

The scalar rule and linearity rules address this gap. They describe how to differentiate functions that are formed by multiplying a function by a constant, and by adding or subtracting functions. These rules allow us to extend our differentiation techniques from simple powers to a broad class of functions constructed from them.

Example 1.11

- If $f(x) = 3x^2$ then $f'(x) = 3 \times \frac{d}{dx}x^2 = 6x$.
- If $g(t) = 3t^2 + 2t^{-2}$ then $g'(t) = \frac{d}{dt}3t^2 + \frac{d}{dt}2t^{-2} = 6t - 4t^{-3}$.
- If $y = \frac{3}{\sqrt{x}} - 2x\sqrt[3]{x} = 3x^{-\frac{1}{2}} - 2x^{\frac{4}{3}}$ then $\frac{dy}{dx} = -\frac{3}{2}x^{-\frac{3}{2}} - \frac{8}{3}x^{\frac{1}{3}}$.
- If $y = -0.3x^{-0.4}$ then $\frac{dy}{dx} = 0.12x^{-1.4}$.
- $\frac{d}{dx}2x^{0.3} = 0.6x^{-0.7}$.

Caution. Although the linearity rule states that $\dfrac{d}{dx}(f(x) \pm g(x)) = f'(x) \pm g'(x)$, this property does *not* extend to products or quotients. In general,

$$\frac{d}{dx}(f(x)g(x)) \neq f'(x)g'(x), \qquad \frac{d}{dx}\left(\frac{f(x)}{g(x)}\right) \neq \frac{f'(x)}{g'(x)}.$$

To differentiate $f(x)g(x)$ or $\frac{f(x)}{g(x)}$ we cannot simply differentiate the factors and then multiply or divide the results. The correct techniques are the *product rule* and the *quotient rule*, which are developed in the next section.

The Product and Quotient Rules

Another common way of combining functions is to multiply them, thereby forming a *product*. The product rule provides the method for differentiating functions constructed in this way.

On the other hand the quotient rule allows us to differentiate functions which are formed by dividing one function by another, i.e. by forming quotients of functions. The quotient rule provides the method for differentiating functions constructed in this way.

Theorem 1.3 (Product and Quotient Rules)

5. **Product Rule:** If $f(x)$ and $g(x)$ are differentiable, then
$$\frac{d}{dx}(f(x)g(x)) = f'(x)g(x) + f(x)g'(x).$$
6. **Quotient Rule:** If $f(x)$ and $g(x)$ are differentiable and $g(x) \neq 0$, then
$$\frac{d}{dx}\left(\frac{f(x)}{g(x)}\right) = \frac{f'(x)g(x) - f(x)g'(x)}{[g(x)]^2}.$$

♡

Example 1.12 Product Rule

- If $y = (x+2)\left(x^2+3\right)$ then $y' = (x+2)2x + 1\left(x^2+3\right)$.
- If $f(x) = \sqrt{x}\left(x^3 - 3x^2 + 7\right)$ then $f'(x) = \sqrt{x}\left(3x^2 - 6x\right) + \frac{1}{2}x^{-\frac{1}{2}}\left(x^3 - 3x^2 + 7\right)$.
- If $z = \left(t^2+3\right)\left(\sqrt{t} + t^3\right)$ then $\frac{dz}{dt} = \left(t^2+3\right)\left(\frac{1}{2}t^{-\frac{1}{2}} + 3t^2\right) + 2t\left(\sqrt{t} + t^3\right)$.

Example 1.13 Quotient Rule

- If $y = \frac{2x^2+3x}{x^3+1}$, then $\frac{dy}{dx} = \frac{\left(x^3+1\right)(4x+3) - \left(2x^2+3x\right)3x^2}{\left(x^3+1\right)^2}$.
- If $g(t) = \frac{t^2+3t+1}{\sqrt{t}+1}$ then $g'(t) = \frac{(\sqrt{t}+1)(2t+3) - \left(t^2+3t+1\right)\left(\frac{1}{2}t^{-\frac{1}{2}}\right)}{(\sqrt{t}+1)^2}$.

In the quotient rule, because of the minus sign in the numerator (i.e. in the top line) it is important to get the terms in the numerator in the correct order. This is often a source of mistakes, so be careful. Decide on your own way of remembering the correct order of the terms.

The Chain Rule

The chain rule is arguably the most important differentiation rule, especially in machine learning. It tells us how to find the derivative of a composite function — a function nested inside another. In software, we constantly compose functions, and the chain rule is the mathematical tool for analyzing how changes propagate through these compositions.

Theorem 1.4 (Chain Rule)

1. If $h(x) = f(g(x))$ is a composite function, then its derivative is the derivative of the outer function (evaluated at the inner function) multiplied by the derivative of the inner function.
$$h'(x) = f'(g(x)) \cdot g'(x).$$

♡

Example 1.14

Find the derivative of $h(x) = (x^3 + 2x)^5$.

Solution: This is a composition where the outer function is $f(u) = u^5$ and the inner function is $g(x) = x^3+2x$. Their derivatives are $f'(u) = 5u^4$ and $g'(x) = 3x^2 + 2$. Applying the chain rule:

$$\begin{aligned} h'(x) &= f'(g(x)) \cdot g'(x) \\ &= 5(x^3 + 2x)^4 \cdot (3x^2 + 2) \end{aligned}$$

◀

Example 1.15

Differentiate $(3x^2 - 5)^3$.

Solution: The first step is always to identify that the expression represents a composite function and then to separate it into its outer and inner components. In this example, the outer function is $(\,\cdot\,)^3$, whose derivative is $3(\,\cdot\,)^2$, and the inner function is $3x^2 - 5$, whose derivative is $6x$. Applying the composite function rule therefore gives

$$\frac{d}{dx}(3x^2 - 5)^3 = 3(3x^2 - 5)^2 \times 6x = 18x(3x^2 - 5)^2.$$

Alternatively, we may introduce the substitution $u = 3x^2 - 5$ and write $y = u^3$. Then

$$\frac{dy}{dx} = \frac{dy}{du} \times \frac{du}{dx} = 3u^2 \times 6x = 18x(3x^2 - 5)^2.$$

◀

Example 1.16

Find $\frac{dy}{dx}$ if $y = \sqrt{x^2 + 1}$.

Solution: The outer function is $\sqrt{\cdot}$, whose derivative is $\frac{1}{2\sqrt{\cdot}}$, and the inner function is $x^2 + 1$, whose derivative is $2x$. Applying the chain rule therefore gives

$$\frac{dy}{dx} = \frac{1}{2\sqrt{x^2 + 1}} \times 2x = \frac{x}{\sqrt{x^2 + 1}}.$$

◀

The chain rule is the core mechanism behind the **backpropagation** algorithm used to train neural networks. An error at the output of a network is a composition of many nested functions (the layers). Backpropagation uses the chain rule repeatedly to calculate the derivative of the error with respect to each weight in the network, telling us how to adjust each weight to improve the model.

Example 1.17 A Multi-Rule Problem

Find the derivative of $h(x) = \frac{x^2}{(3x+1)^4}$.

Solution: This problem requires the quotient rule, and the denominator requires the chain rule. Let $f(x) = x^2$ and $g(x) = (3x + 1)^4$. Then $f'(x) = 2x$. To find $g'(x)$, we use the chain rule: $g'(x) = 4(3x + 1)^3 \cdot 3 =$

$12(3x+1)^3$. Now, apply the quotient rule:

$$\begin{aligned} h'(x) &= \frac{f'(x)g(x) - f(x)g'(x)}{[g(x)]^2} \\ &= \frac{(2x)(3x+1)^4 - (x^2)(12(3x+1)^3)}{((3x+1)^4)^2} \\ &= \frac{(2x)(3x+1) - 12x^2}{(3x+1)^5} \quad \text{(after factoring out and canceling } (3x+1)^3\text{)} \\ &= \frac{6x^2 + 2x - 12x^2}{(3x+1)^5} = \frac{2x - 6x^2}{(3x+1)^5} \end{aligned}$$

◀

Summary of Rules of Differentiation

The rules of differentiation are summarized in the following theorem.

Theorem 1.5 (Rules of Differentiation)

1. **Constant Rule:** If c is constant, then $\frac{d}{dx}c = 0$.
2. **Power Rule:** For any real n, $\frac{d}{dx}x^n = nx^{n-1}$.

   $$\frac{d}{dx}x^n = nx^{n-1}.$$
3. **Scalar Rule:** If c is constant, then $\frac{d}{dx}[c \cdot f(x)] = c \cdot f'(x)$.
4. **Linearity:** For differentiable f, g and constants a, b, $\frac{d}{dx}(af + bg) = af' + bg'$.
5. **Product Rule:** If f, g are differentiable, then $\frac{d}{dx}(fg) = f'g + fg'$.
6. **Quotient Rule:** If f, g are differentiable and $g(x) \neq 0$, then $\frac{d}{dx}\left(\frac{f}{g}\right) = \frac{f'g - fg'}{g^2}$.
7. **Chain Rule:** If f, g are differentiable, then $\frac{d}{dx}(f(g(x))) = f'(g(x))g'(x)$.

♡

These rules do not apply to every function. For instance, the derivative of e^x is e^x, while the derivative of $\ln x$ is $\frac{1}{x}$, and neither of these fits the patterns described so far. The most common differentiable functions that fall outside these earlier rules are listed below.

- $\frac{d}{dx}e^x = e^x$
- $\frac{d}{dx}\ln x = \frac{1}{x}$
- $\frac{d}{dx}\sin x = \cos x$
- $\frac{d}{dx}\cos x = -\sin x$

Example 1.18

Differentiate $\ln(x^2 + 3x + 1)$.

Solution: We solve this by using the chain rule and our knowledge of the derivative of $\ln x$.

$$\begin{aligned}
\frac{d}{dx}\ln\left(x^2+3x+1\right) &= \frac{d}{dx}\left(\ln u\right) \quad \left(\text{ where } u = x^2+3x+1\right)\\
&= \frac{d}{du}\left(\ln u\right)\times\frac{du}{dx} \quad (\text{ by the chain rule })\\
&= \frac{1}{u}\times\frac{du}{dx}\\
&= \frac{1}{x^2+3x+1}\times\frac{d}{dx}\left(x^2+3x+1\right)\\
&= \frac{1}{x^2+3x+1}\times(2x+3)\\
&= \frac{2x+3}{x^2+3x+1}
\end{aligned}$$

◀

Example 1.19

Find $\frac{d}{dx}\left(e^{3x^2}\right)$.

Solution: This is an application of the chain rule together with our knowledge of the derivative of e^x.

$$\begin{aligned}
\frac{d}{dx}\left(e^{3x^2}\right) &= \frac{de^u}{dx} \quad \text{where } u = 3x^2\\
&= \frac{de^u}{du}\times\frac{du}{dx} \quad \text{by the chain rule}\\
&= e^u\times\frac{du}{dx}\\
&= e^{3x^2}\times\frac{d}{dx}\left(3x^2\right)\\
&= 6xe^{3x^2}
\end{aligned}$$

◀

Example 1.20

Find $\frac{d}{dx}\left(e^{x^2+2x}\right)$.

Solution: Again, we use our knowledge of the derivative of e^x together with the chain rule.

$$\begin{aligned}
\frac{d}{dx}\left(e^{x^2+2x}\right) &= \frac{de^{\mathrm{u}}}{dx} \quad \left(\text{ where } u = x^3+2x\right)\\
&= e^{\mathrm{z}}\times\frac{du}{dx} \quad \text{(by the chain rule)}\\
&= e^{x^3+2x}\times\frac{d}{dx}\left(x^3+2x\right)\\
&= \left(3x^2+2\right)\times e^{x^2+2x}
\end{aligned}$$

◀

Example 1.21

Differentiate $\ln\left(2x^3+5x^2-3\right)$.

Solution: We solve this by using the chain rule and our knowledge of the derivative of $\ln x$.

$$\begin{aligned}
\frac{d}{dx}\ln\left(2x^3+5x^2-3\right) &= \frac{d\ln u}{dx} \quad \text{(where } u=\left(2x^3+5x^2-3\right)\\
&= \frac{d\ln u}{du}\times\frac{du}{dx} \quad \text{(by the chain rule)}\\
&= \frac{1}{u}\times\frac{du}{dx}\\
&= \frac{1}{2x^3+5x^2-3}\times\frac{d}{dx}\left(2x^3+5x^2-3\right)\\
&= \frac{1}{2x^3+5x^2-3}\times\left(6x^2+10x\right)\\
&= \frac{6x^2+10x}{2x^3+5x^2-3}
\end{aligned}$$

◀

There are two shortcuts to differentiating functions involving exponents and logarithms. The four examples above gave

$$\begin{aligned}
\frac{d}{dx}\left(\ln\left(x^2+3x+1\right)\right) &= \frac{2x+3}{x^2+3x+1}\\
\frac{d}{dx}\left(e^{3x^2}\right) &= 6xe^{3x^2}\\
\frac{d}{dx}\left(e^{x^2+2x}\right) &= \left(3x^2+2\right)e^{x^2+2x}\\
\frac{d}{dx}\left(\ln\left(2x^3+5x^2-3\right)\right) &= \frac{6x^2+10x}{2x^3+5x^2-3}
\end{aligned}$$

These examples suggest the general rules

$$\begin{aligned}
\frac{d}{dx}\left(e^{f(x)}\right) &= f'(x)e^{f(x)}\\
\frac{d}{dx}(\ln f(x)) &= \frac{f'(x)}{f(x)}
\end{aligned}$$

These rules arise from the chain rule and the fact that $\frac{dx^x}{dx}=e^x$ and $\frac{d\ln x}{dx}=\frac{1}{x}$. They can speed up the process of differentiation but it is not necessary that you remember them. If you forget, just use the chain rule as in the examples above.

1.4 Applications of Derivatives

The development of mathematics ranks among the greatest achievements of human thought, and the emergence of calculus — both differential and integral — marks a turning point in that history. Differential calculus in particular has countless practical applications across science and engineering, far too many to enumerate. Its importance is reflected in the fact that virtually every quantitative discipline relies on its methods.

Within elementary mathematics, differential calculus is used primarily in two areas: analysing and sketching curves, and solving optimisation problems. In this section, we offer a brief introduction to how differential calculus is applied in optimisation.

Stationary points - the idea behind optimisation

Let us return to our metaphor of the mountain range. Imagine standing somewhere among peaks and valleys, perhaps even with your eyes closed or shrouded in fog. As you wander along, you notice that whenever you're hiking uphill or downhill, you know you are not at the highest summit or the lowest valley—you are either still climbing or still descending.

But at the top of a peak, there is a brief, level stretch where the slope under your feet is zero. At a bottom of a valley, there is also a level spot. If you find yourself on perfectly flat ground—at least for a moment—you have a clue: you might be at a summit, a valley, or perhaps at a gentle inflection point along a ridge. In the vast terrain of a mountain range, it is only at these level patches, where the slope vanishes, that you could possibly be at an extreme point: a maximum or minimum.

This is the key insight behind optimisation in calculus: to find the highest or lowest points (the peaks or valleys) of a function—in other words, to solve optimisation problems—we need only to examine the locations where the "slope" is zero. These locations are known as stationary points.

Definition 1.4

For a function $y = f(x)$ the points on the graph where the graph has zero slope are called stationary points. In other words stationary points are where $f'(x) = 0$.

To find the stationary points of a function we differentiate, set the derivative equal to zero and solve the equation.

Example 1.22

Find the stationary points of the function $f(x) = 2x^3 + 3x^2 - 12x + 17$.

Solution:

$f'(x) = 6x^2 + 6x - 12$. Setting $f'(x) = 0$ and solving we obtain

$$\begin{aligned} 6x^2 + 6x - 12 &= 0 \\ x^2 + x - 2 &= 0 \\ (x-1)(x+2) &= 0 \\ x &= 1, -2 \end{aligned}$$

This gives us the values of x for which the function f is stationary. The corresponding values of the function are found by substituting 1 and -2 into the function.

They are $f(1) = 2\times1^3+3\times1^2-12\times1+17 = 10$ and $f(-2) = 2\times(-2)^3+3\times(-2)^2-12\times(-2)+17 = 37$. The stationary points are therefore $(1, 10)$ and $(-2, 37)$. ◀

Example 1.23

Find the stationary points of the function $g(t) = e^{t^2}$.

Solution: Differentiating and setting the derivative equal to zero we obtain the equation $g'(t) = 2te^{t^2} = 0$. Since e^{t^2} is never zero, the only solution to this equation is where $2t = 0$, ie $t = 0$. Substituting into the formula for g we obtain the function value $g(0) = e^{0^2} = 1$. Thus the stationary point is $(0, 1)$. ◀

Types of stationary points

Returning to our mountain range metaphor, stationary points can be thought of as the special spots where, just for a moment, the ground feels perfectly level beneath your feet. Sometimes, this happens at the very top of a hill—where, if you reach out in any direction, you only sense downhill slopes ahead. This is called a local maximum: the summit of your immediate surroundings, though not necessarily the tallest peak in the whole range. Other times, the level spot is at the bottom of a valley—everything around you slopes upward. This is called a local minimum: the lowest place in your local vicinity, even if deeper valleys exist elsewhere. The "local" label emphasizes that we're talking about the highest or lowest point nearby, not globally. You might also hear the terms "relative maximum" and "relative minimum" used for these. Figure 1.5 illustrates a function as a landscape with both a hilltop (local maximum) and a valley floor (local minimum). Notice that, at each of these points—just like on a level patch in the mountains—the slope is zero.

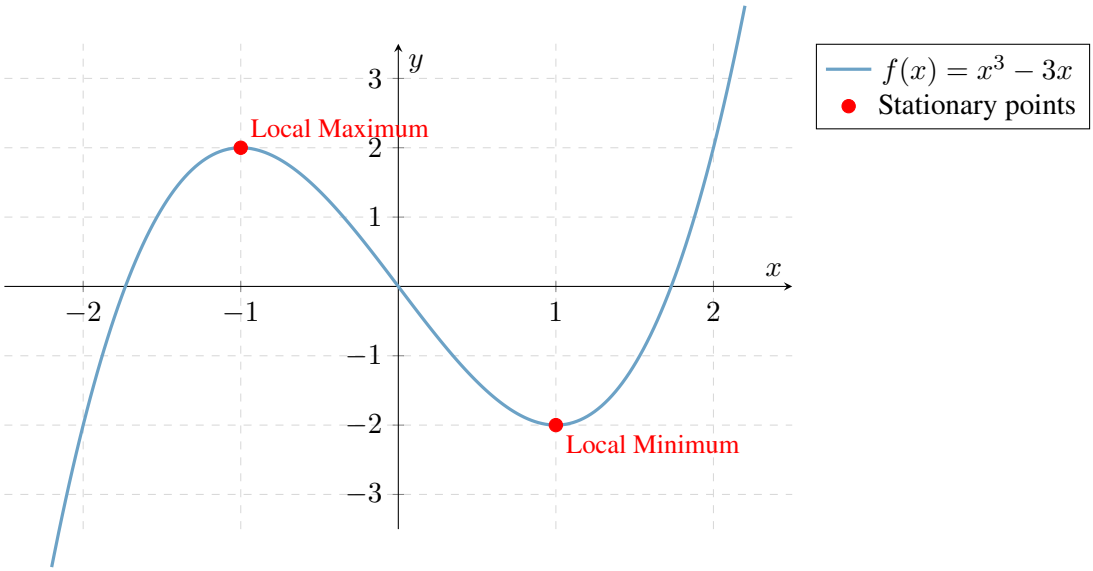


Figure 1.5: A curve with a local maximum at $x = -1$ and a local minimum at $x = 1$.

Local maxima and local minima are not the only types of stationary points. There is a third kind. Figure 1.6 shows a stationary point that is neither a local maximum nor a local minimum. This type of stationary point is called a stationary point of inflection or simply *inflection point*.

The first derivative test

Let us return to Example 1.22. For the function $f(x) = 2x^3 + 3x^2 - 12x + 17$ we identified stationary points at $(1, 10)$ and $(-2, 37)$. The natural question is: what type of stationary points are these? Without a sketch of the graph it is not immediately clear whether they correspond to local maxima, local minima, or stationary points of inflection. Drawing an accurate graph would resolve the question, but doing so can require substantial effort. Instead, we seek a method that allows us to classify a stationary point without needing to plot the entire function.

Several approaches exist, but in this section we focus on one technique, known as the *first derivative test*. The idea is to examine the behaviour of the function immediately to the left and immediately to the right of the stationary point.

To understand the principle, imagine a person standing at a point on a landscape where the ground is level.

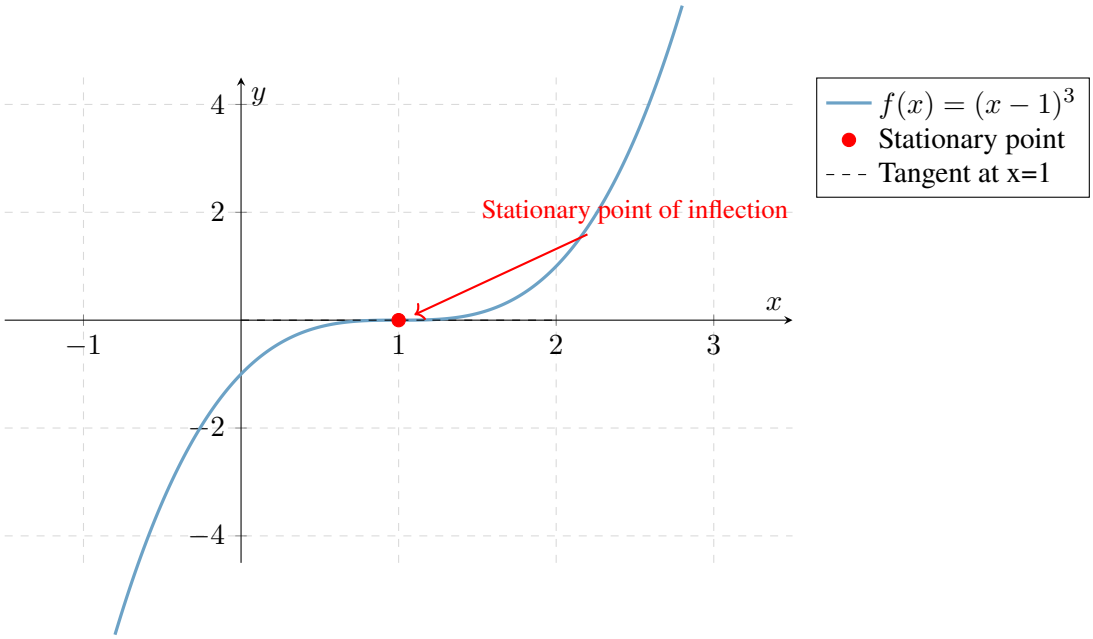


Figure 1.6: A function with a stationary point of inflection at $x = 1$. The tangent line at this point is horizontal, but it is neither a local maximum nor a local minimum.

Without seeing the surroundings, the person wishes to determine whether this point is the top of a hill, the bottom of a valley, or neither. One way to decide is to take a small step backward and observe the slope, and then take a small step forward and observe the slope again.

If the ground slopes upward behind and downward ahead, the level point must be the top of a hill, corresponding to a local maximum. If the ground slopes downward behind and upward ahead, the point must be the bottom of a valley, corresponding to a local minimum. The remaining possibility-where the slopes have the same sign on both sides-indicates a stationary point of inflection.

This reasoning captures the essence of the first derivative test: by analysing the sign of $f'(x)$ immediately on either side of a stationary point, we can determine its character without relying on a graph.

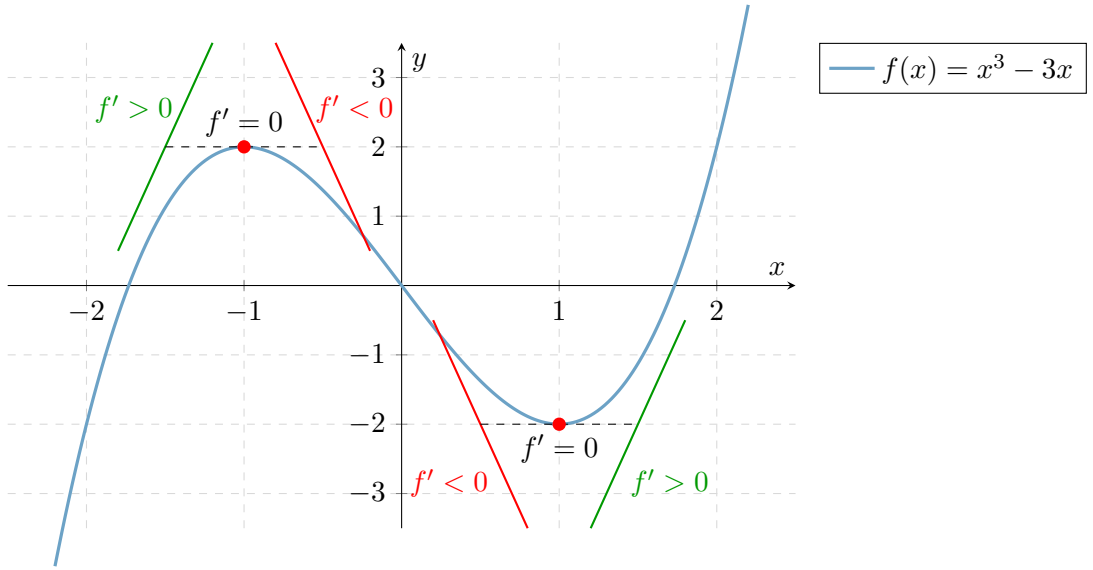


Figure 1.7: Illustration of the first derivative test. At a local maximum, the derivative changes from positive to negative. At a local minimum, it changes from negative to positive.

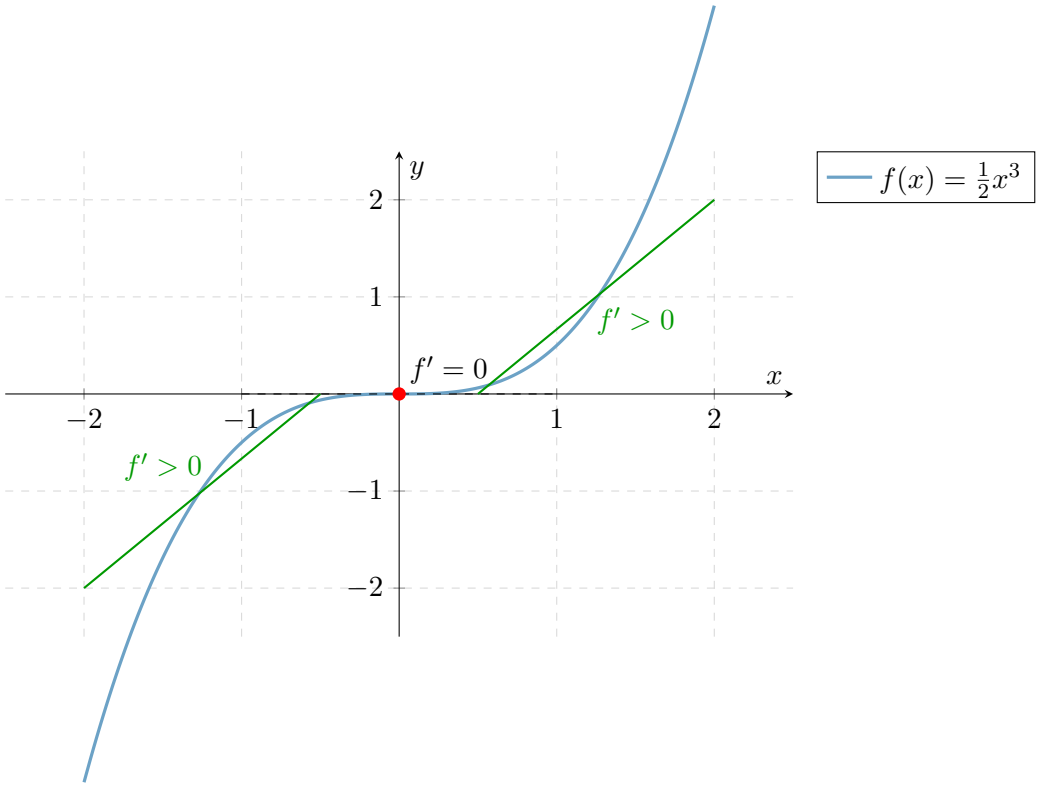


Figure 1.8: Illustration of a stationary point of inflection. The derivative is zero at $x = 0$, but it is positive on both sides, so the point is neither a local maximum nor a minimum.

We can summarize these considerations in the following theorem:

Theorem 1.6 (The First Derivative Test)

Let c be a critical point of a continuous function f.

- If $f'(x)$ changes from negative to positive at c, then f has a **local minimum** at c.
- If $f'(x)$ changes from positive to negative at c, then f has a **local maximum** at c.
- If $f'(x)$ does not change sign at c, then f has no local extremum at c; it is a stationary point of inflection.

♡

This test is visualized in Figure 1.7 and Figure 1.8.

Optimisation

Now that we've laid the groundwork, we can approach some optimisation problems. When we want to maximise a function $f(x)$ within a specified interval for x, our goal is to find the largest value that $f(x)$ achieves on that interval. Importantly, this maximum value does not always coincide with a stationary point. As shown in Figure 1.9, consider searching for the greatest and least values of a function on the interval $2 \leq x \leq 7$. Within this interval, the function has two stationary points—one corresponding to a local maximum, the other to a local minimum. However, the highest value of the function in this interval actually occurs at the endpoint $x = 7$, which is not a stationary point. It takes this value simply because, outside of the interval, larger x values are excluded from consideration. In contrast, the minimum value in this example is located at a stationary point within the interval. With this understanding, we can now explain the systematic steps for finding the maximum or minimum of a function in a given region.

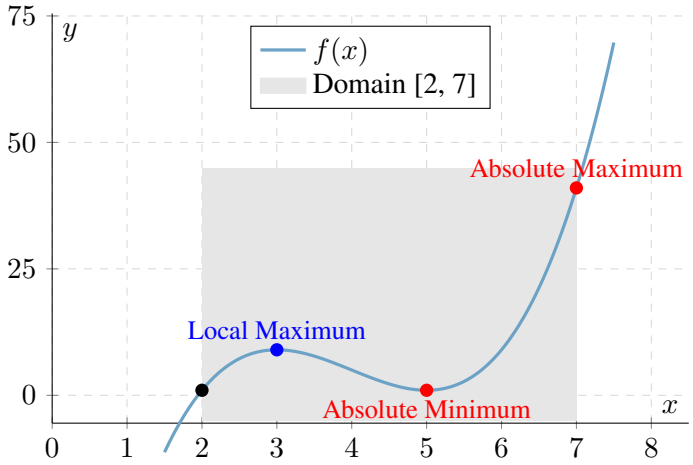


Figure 1.9: On the restricted domain $[2, 7]$, the absolute maximum occurs at an endpoint ($x = 7$), not at the local maximum. The absolute minimum occurs at an interior stationary point ($x = 5$).

The location of maxima and minima

A function $f(x)$ may or may not have a maximum or minimum value in a particular region of x values. However, if they do exist the maximum and the minimum values must occur at one of three places:

1. At the endpoints (if they exist) of the region under consideration.
2. Inside the region at a stationary point.
3. Inside the region at a point where the derivative does not exist.

Remark:

1. It is straightforward to find cases where a function does not attain a maximum or minimum within a certain range. For instance, the function $f(x) = x$ lacks both a maximum and a minimum over the interval $-\infty < x < \infty$; its graph increases without bound as x increases. According to Point 1 above, note that in the region $-\infty < x < \infty$, there are actually no endpoints. As another illustration, for the region $x \geq 1$, there is only a single endpoint at $x = 1$.
2. Regarding Point 3, this text does not cover situations where the derivative fails to exist at certain points. Nevertheless, keep in mind that such points can occur and could be where a maximum or minimum is located. For more details, refer to advanced calculus resources.

Now that we know exactly where the maxima or minima can occur, we can give a procedure for finding them.

Procedure for finding the maximum or minimum values of a function.

1. Find the endpoints of the region under consideration (if there are any).
2. Find all the stationary points in the region.
3. Find all points in the region where the derivative does not exist.
4. Substitute each of these into the function and see which gives the greatest (or smallest) function value.

Example 1.24

Find the minimum value and the maximum value of the function $f(x) = x^2 e^x$ for $-4 \leq x \leq 1$.

Solution: We will follow the procedure outlined above. The endpoints are -4 and 1 . Differentiating we obtain $f'(x) = x^2e^x + 2xe^x = x(x+2)e^x$. Setting $f'(x) = 0$ and solving we get stationary points at $x = 0$ and $x = -2$. There are no points where the derivative does not exist. Therefore the maximum and minimum values will be found at one of the points $x = -4, -2, 0, 1$. Substituting we obtain $f(-4) \approx 0.29, f(-2) \approx 0.54, f(0) = 0$ and $f(1) = e \approx 2.7$. therefore the maximum value occurs at $x = 1$ and is equal to e, and the minimum value occurs at $x = 0$ and is 0 . ◀

Example 1.25

Find the maximum and minimum values of the function $g(t) = \frac{1}{3}t^3 - t + 2$ for $0 \leq t \leq 3$.

Solution: The endpoints are $t = 0$ and $t = 3$. Differentiating and equating to zero we get $g'(t) = t^2 - 1 = (t-1)(t+1) = 0$ so the stationary points are at $t = -1, 1$. Since -1 is not in the region, the possible locations of the maximum and the minimum are $t = 0, 1, 3$. Substituting into g we obtain $g(0) = 2, g(1) = \frac{4}{3}$ and $g(3) = 8$. The maximum is therefore $g(3) = 8$ and the minimum is $g(1) = \frac{4}{3}$. ◀

Example 1.26

Suppose you are tuning a machine learning model, and you want to find the learning rate α (between 0 and 1) that minimizes the loss function $L(\alpha) = (\alpha - 0.3)^2 + 1$. What value of α gives the minimum loss?

Solution: We will apply the procedure for finding absolute extrema on the closed interval $[0, 1]$. The function to minimize is the loss function $L(\alpha) = (\alpha - 0.3)^2 + 1$.

1. **Identify Endpoints:** The region under consideration is $0 \leq \alpha \leq 1$, so the endpoints are $\alpha = 0$ and $\alpha = 1$.
2. **Find Stationary Points:** We first find the derivative of the loss function with respect to α:
   $$\frac{dL}{d\alpha} = \frac{d}{d\alpha}\left[(\alpha - 0.3)^2 + 1\right] = 2(\alpha - 0.3)$$
   Next, we set the derivative to zero to find the stationary points:
   $$2(\alpha - 0.3) = 0 \implies \alpha = 0.3$$
   The only stationary point is $\alpha = 0.3$, which lies within our interval $[0, 1]$.
3. **Identify Points Where Derivative is Undefined:** The derivative $2(\alpha - 0.3)$ is a simple linear function and is defined for all values of α. There are no such points.
4. **Compare Values at Candidate Points:** We now evaluate the loss function $L(\alpha)$ at all the candidate points we have found: the endpoints and the stationary point.
   - At the left endpoint, $\alpha = 0$:
     $$L(0) = (0 - 0.3)^2 + 1 = 0.09 + 1 = 1.09$$
   - At the stationary point, $\alpha = 0.3$:
     $$L(0.3) = (0.3 - 0.3)^2 + 1 = 0 + 1 = 1.00$$
   - At the right endpoint, $\alpha = 1$:
     $$L(1) = (1 - 0.3)^2 + 1 = (0.7)^2 + 1 = 0.49 + 1 = 1.49$$

By comparing these values, we see that the smallest loss is 1.00, which occurs at the stationary point.

Conclusion: The minimum loss occurs when the learning rate is $\alpha = 0.3$. ◀

1.5 The Second Derivative: Curvature and Inflection

The first derivative, $f'(x)$, tells us about the slope or rate of change of a function. The **second derivative**, denoted $f''(x)$, is the derivative of the first derivative. It measures how the slope itself is changing. This provides deeper insight into the shape of a function's graph.

Definition 1.5 (The Second Derivative)

The **second derivative** of a function f, denoted $f''(x)$, is the function

$$f''(x) = \frac{d}{dx}\left(f'(x)\right)$$

provided the limit exists.

The most important geometric interpretation of the second derivative is **concavity**.

- If $f''(x) > 0$ on an interval, the slope $f'(x)$ is increasing. The graph bends upwards, like a cup holding water. We say the function is **concave up**.
- If $f''(x) < 0$ on an interval, the slope $f'(x)$ is decreasing. The graph bends downwards, like a cup spilling water. We say the function is **concave down**.

A point on the graph where the concavity changes is called an **inflection point**. This typically occurs where $f''(x) = 0$.

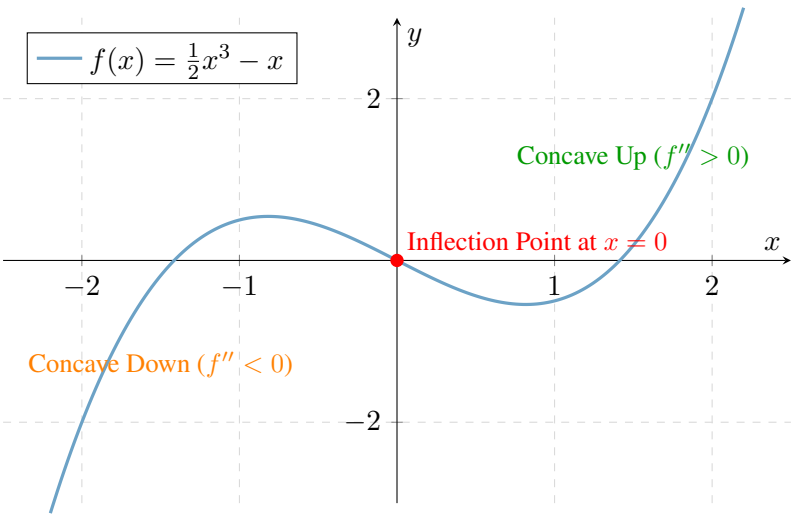


Figure 1.10: A function showing a region of concave down curvature, followed by a region of concave up curvature. The transition occurs at the inflection point.

The second derivative also gives us a powerful tool for classifying stationary points.

Theorem 1.7 (The Second Derivative Test)

Let c be a stationary point of f (i.e., $f'(c) = 0$).

- If $f''(c) > 0$, then f has a **local minimum** at c.
- If $f''(c) < 0$, then f has a **local maximum** at c.
- If $f''(c) = 0$, the test is inconclusive.

Chapter 2 Multivariable Calculus and Gradients

In the previous chapter, we explored the derivative as a tool for measuring the rate of change of a function of a single variable. Geometrically, this corresponded to finding the slope of a curve at a point. However, many real-world systems depend on multiple factors. The fuel efficiency of an aircraft depends on both its speed and altitude. The performance of a machine learning model is a function of thousands, or even millions, of parameters. To navigate these complex relationships, we must extend our understanding of calculus to higher dimensions.

This brings us from the simple path of a curve to the vast terrain of a surface. Imagine again our hiker, but now standing on a mountain range. The "slope" is no longer a single number; it depends entirely on the direction the hiker chooses to face. Pointing straight up the mountain reveals the steepest path, while facing along the mountain's contour results in a perfectly level path. The mathematical tool that captures this multidimensional slope is the **gradient**.

2.1 Functions of Several Variables

We now extend our view from functions of a single variable, $f(x)$, to functions of two or more variables. For simplicity, we will focus on functions of two variables, written as $z = f(x, y)$. The main difference is thus that, instead of mapping values of one variable to values of another variable, we map ordered pairs of variables to another variable.

Definition 2.1 (Function of Two Variables)

A *function of two variables* is a rule that assigns to each ordered pair (x, y) in a set $D \subseteq \mathbb{R}^2$ a unique real number z, which we write as $z = f(x, y)$. The set D is called the *domain* of f.
The *range* of f is the set of all real numbers z for which there exists at least one $(x, y) \in D$ such that $f(x, y) = z$ which is illustrated in Figure 2.1.

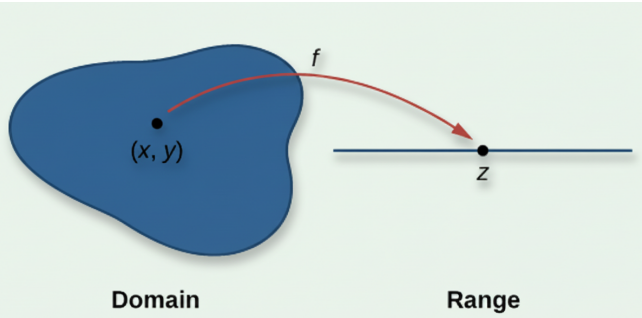


Figure 2.1: The domain of a function of two variables consists of ordered pairs (x, y).

Example 2.1 Find the domain and range of the function $f(x, y) = \sqrt{9 - x^2 - y^2}$.

Solution: To determine the domain, we require the expression inside the square root to be non-negative:

$$9 - x^2 - y^2 \geq 0.$$

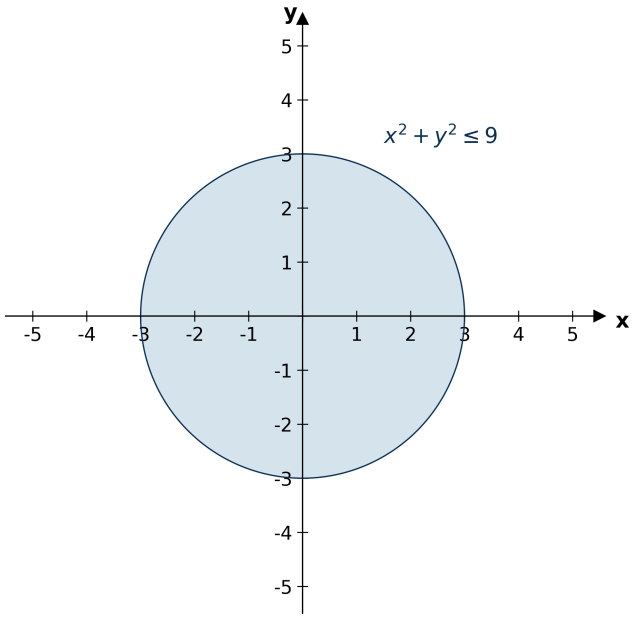


Figure 2.2: The domain of the function $g(x, y) = \sqrt{9 - x^2 - y^2}$ is a closed disk of radius 3 .

Rearranging gives the inequality

$$x^2 + y^2 \leq 9.$$

Therefore, the domain of f is the closed disk of radius 3 centred at the origin, which is illustrated in Figure 2.2, and explicitly given by

$$\mathrm{Dom}(f) = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 9\}.$$

To find the range, note that for any point (x, y) in the domain we have

$$0 \leq x^2 + y^2 \leq 9,$$

so

$$0 \leq 9 - x^2 - y^2 \leq 9.$$

Taking square roots (and remembering the square root is always non-negative) gives

$$0 \leq f(x, y) = \sqrt{9 - x^2 - y^2} \leq 3.$$

Both endpoints of this interval are attainable:

- $f(x, y) = 3$ at the centre $(0, 0)$,
- $f(x, y) = 0$ on the boundary circle $x^2 + y^2 = 9$.

Thus, the range of f is the closed interval

$$\mathrm{Ran}(f) = \{z \in \mathbb{R} \mid 0 \leq z \leq 3\} = [0, 3].$$

◀

While the graph of $f(x)$ is a curve in a 2D plane, the graph of $f(x, y)$ is a **surface** in 3D space. When graphing a function of two variables $z = f(x, y)$, each point (x, y) in the domain is assigned a height z, producing an ordered triple (x, y, z). The collection of all such triples forms a *surface* in three-dimensional space. One can

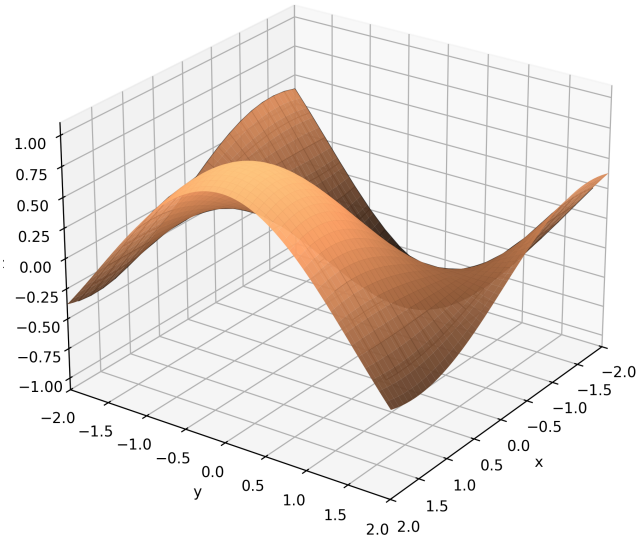


Figure 2.3: An example of a function of two variables, ploted in 3D.

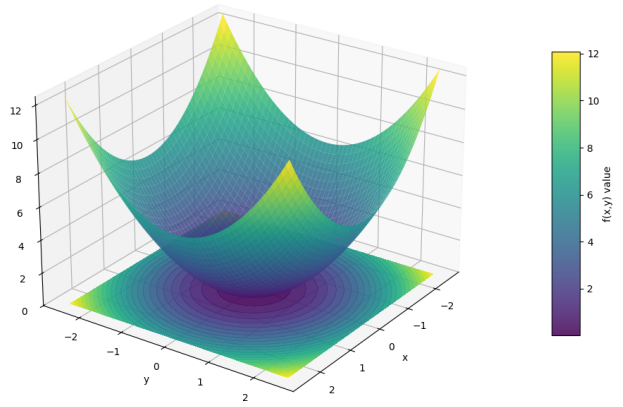
think of the xy-plane lying flat, with the value of z plotted vertically above (or below) each point (x, y). The resulting surface illustrates how the function behaves across its domain, as shown in Figure 2.3.

Definition 2.2 (Level Curve and Contour Map)

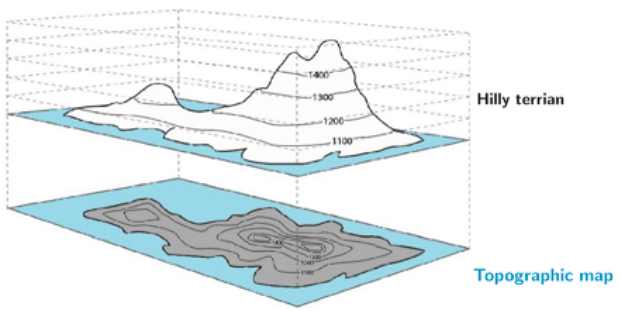
A **level curve** of a function $f(x, y)$ is the set of all points (x, y) in the input plane where the function has a constant value, i.e., $f(x, y) = c$ for some constant c.

A collection of level curves for different values of c forms a **contour map**.

Figure 2.4: Contour Maps: (a) a 3D surface plot with contour lines projected onto the xy-plane, and (b) a topographic map showing level curves.



(a) The surface $z = x^2 + y^2$ with contour lines projected onto the xy-plane. Each contour corresponds to a constant function value, illustrating how level curves arise from horizontal slices of the surface.



(b) A topographic map of a hilly terrain.

On a geographical map, these lines represent paths of constant elevation. On our function surface, they represent paths of constant "height" z. Figure 2.4a shows the surface $z = x^2 + y^2$ with contour lines projected onto

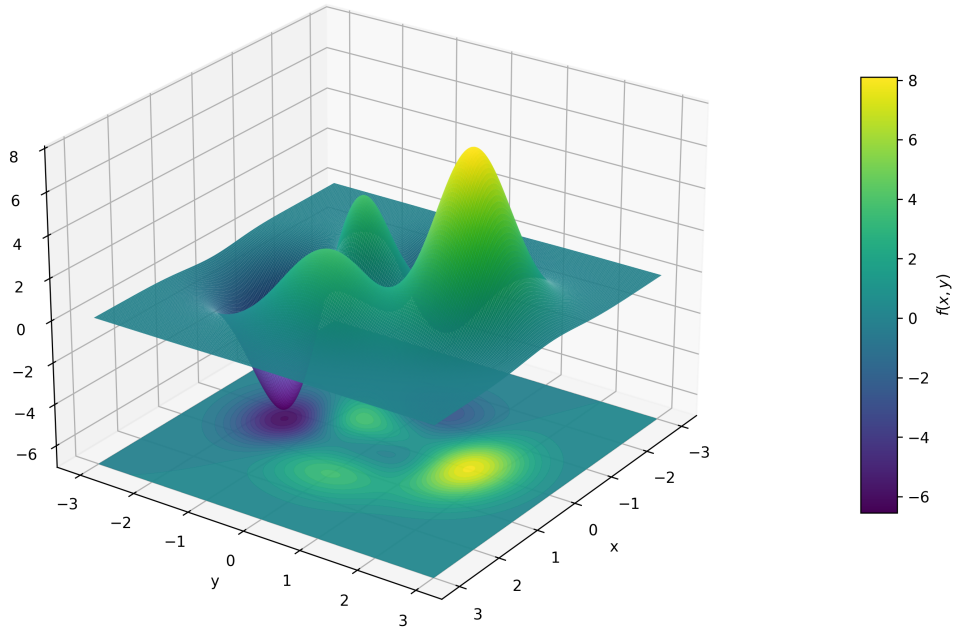


Figure 2.5: A more complex function of two variables, plotted in 3D with coloured contour lines projected onto the xy-plane.

the xy-plane. Each contour corresponds to a constant function value, illustrating how level curves arise from horizontal slices of the surface. Figure 2.4b shows a topographic map of a hilly terrain. These contour lines are not limited to simple functions; they can also represent much more intricate surfaces, such as the one shown in Figure 2.3.

2.2 Partial Derivatives and the Gradient

How do we measure the rate of change of a multivariable function? Since the slope depends on the direction, we can start with the simplest directions: parallel to the coordinate axes. This leads to the idea of a **partial derivative**.

Partial Derivatives

For a function $z = f(x, y)$, we analyse how the surface changes by allowing one variable to vary while keeping the other fixed. The resulting rates of change are the *partial derivatives* of f. The partial derivative with respect to x is computed by treating y as constant and differentiating with respect to x; the derivative with respect to y is obtained similarly. These two quantities, $\partial z/\partial x$ and $\partial z/\partial y$, describe how steeply the surface rises or falls in the coordinate directions.

Geometrically, $\frac{\partial f}{\partial x}$ at a point (a, b) represents the slope of the surface in the direction of the x-axis and $\frac{\partial f}{\partial y}$ represents the slope of the surface in the direction of the y-axis.

Definition 2.3 (Partial Derivatives)

The **partial derivative** of $f(x, y)$ with respect to x, denoted $\frac{\partial f}{\partial x}$ or f_x, is

$$\frac{\partial f}{\partial x} = \lim_{h \to 0} \frac{f(x+h, y) - f(x, y)}{h}$$

The **partial derivative** with respect to y, denoted $\frac{\partial f}{\partial y}$ or f_y, is

$$\frac{\partial f}{\partial y} = \lim_{h \to 0} \frac{f(x, y+h) - f(x, y)}{h}$$

♣

Example 2.2 Consider the function

$$f(x, y) = x^2 + 2y^2.$$

We compute its partial derivatives by differentiating with respect to one variable while holding the other constant.

Partial derivative with respect to x.

Treat y as a constant and differentiate as in single–variable calculus:

$$\frac{\partial}{\partial x}(x^2 + 2y^2) = 2x.$$

Geometrically, this corresponds to slicing the surface $z = f(x, y)$ with a plane $y =$ constant, which produces the parabola $z = x^2 +$ constant. The value $2x$ is the slope of this parabola at the chosen point.

Partial derivative with respect to y.

Now treat x as constant:

$$\frac{\partial}{\partial y}(x^2 + 2y^2) = 4y.$$

This represents the slope of the cross–section obtained by slicing the surface with a plane $x =$ constant.

Example 2.3 Find the partial derivatives of $f(x, y) = 3x^2 + 2xy^3 + 5y$.

Solution:

To find $\frac{\partial f}{\partial x}$, we treat y as a constant:

$$\frac{\partial f}{\partial x} = \frac{\partial}{\partial x}(3x^2) + \frac{\partial}{\partial x}(2xy^3) + \frac{\partial}{\partial x}(5y) = 6x + 2y^3 + 0 = 6x + 2y^3$$

To find $\frac{\partial f}{\partial y}$, we treat x as a constant:

$$\frac{\partial f}{\partial y} = \frac{\partial}{\partial y}(3x^2) + \frac{\partial}{\partial y}(2xy^3) + \frac{\partial}{\partial y}(5y) = 0 + 2x(3y^2) + 5 = 6xy^2 + 5$$

◀

The Gradient Vector

The partial derivatives give us the rate of change in two specific directions. The **gradient** combines this information into a single vector that points in the direction of the *steepest* rate of change.[1] The word "gradient"

[1] This property is formally proven using the concept of the *directional derivative*, which measures the rate of change in any arbitrary direction.

in mathematics describes a vector that indicates both the direction and the steepness of ascent for a function of several variables:

$$\nabla f(x, y) = \left\langle \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \right\rangle$$

The symbol ∇ is called the **nabla** or **del** operator.

Definition 2.4 (The Gradient)

The **gradient** of a function $f(x, y)$, denoted ∇f, is the vector function defined by:

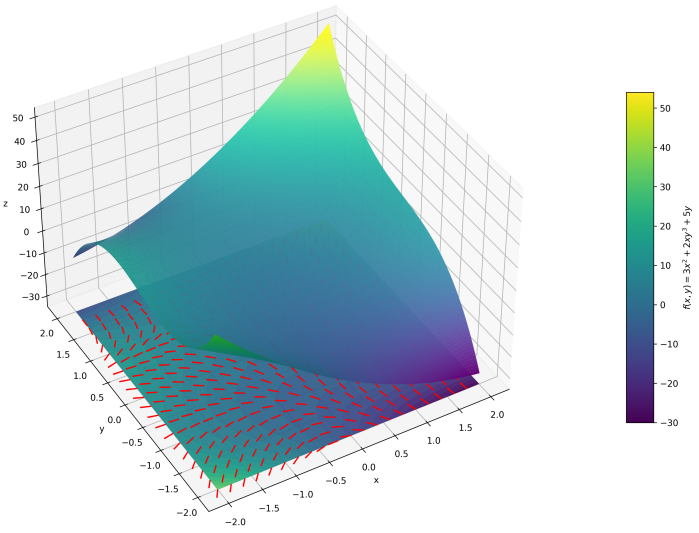
$$\nabla f(x, y) = \left\langle \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \right\rangle = \frac{\partial f}{\partial x}\mathbf{i} + \frac{\partial f}{\partial y}\mathbf{j}$$

The gradient is the multivariable analogue of the first derivative. It has two crucial geometric properties:

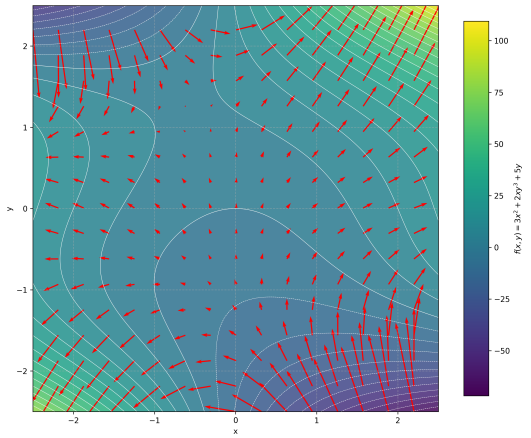
1. **Direction of Steepest Ascent:** At any point (x, y), the gradient vector $\nabla f(x, y)$ points in the direction in which the function f increases most rapidly.
2. **Perpendicularity to Level Curves:** The gradient vector $\nabla f(x, y)$ is always orthogonal (perpendicular) to the level curve of f that passes through the point (x, y).

Both properties are extremely useful. The first tells us that the gradient points in the direction of steepest ascent, while the second explains that this direction is perpendicular to the contour lines. Consequently, to move uphill as fast as possible, one should walk in the direction of the gradient; to move downhill as fast as possible, one should walk in the opposite direction, that is, along the negative gradient.

Figure 2.6: Two perspectives of the function $f(x, y) = 3x^2 + 2xy^3 + 5y$. The 3D view (a) shows the physical surface, while the 2D view (b) shows the contour map we use for analysis. The red gradient vectors point in the direction of steepest ascent and are always perpendicular to the level curves.



(a) A 3D view of the surface $z = f(x, y)$ with its gradient field projected below.



(b) A 2D top-down view showing the contour map and the gradient field.

To fully appreciate the geometric meaning of the gradient, it is helpful to visualize it from two different perspectives, as shown in Figure 2.6. The figure presents the same function, $f(x, y) = 3x^2 + 2xy^3 + 5y$, in

two ways: as a three-dimensional surface and as a two-dimensional contour map.

On the left, in Figure 2.6a, we see the function as a landscape. The height of the surface at any point (x, y) corresponds to the value of $z = f(x, y)$. Below this surface, we have projected its contour map along with its gradient field, represented by red arrows. You can visually trace how the steepness of the surface at a point relates to the arrow directly beneath it.

On the right, in Figure 2.6b, we have the perspective we most often work with: a top-down view of the xy-plane. This is the standard contour map, where each line connects points of equal "elevation." The gradient vectors are overlaid on this map.

By comparing these two views, we can confirm the two fundamental properties of the gradient:

1. **Direction of Steepest Ascent:** Look at any red arrow on the 2D map. Notice that it always points from a lower-value region (lighter colours) towards a higher-value region (darker colours). The arrows show the "uphill" direction. This corresponds directly to the steepest path one could take up the surface in the 3D view.
2. **Perpendicularity to Level Curves:** Observe the relationship between the red arrows and the white contour lines in Figure 2.6b. At every point, the gradient vector is perfectly orthogonal (perpendicular) to the level curve passing through that point. This makes intuitive sense: to climb a hill most efficiently, you must walk perpendicular to the paths of level elevation.

This side-by-side comparison bridges the gap between the physical reality of the function's shape and the mathematical tools we use to analyze it. While in machine learning we cannot visualize a function with a million dimensions, the principle remains the same: we compute the gradient vector to find the direction of steepest ascent and use its negative, $-\nabla f$, to guide us "downhill" towards a minimum.

2.3 Applications of the Gradient

The gradient is one of the most fundamental concepts in applied mathematics, with profound implications in physics, engineering, and computer science, and lies as the cornerstone in recent developments in machine learning and artificial intelligence.

Optimisation and Gradient Descent

Remember our hiker trying to find the lowest point in a foggy valley. The most efficient strategy is to always walk in the direction of steepest *descent*. From the properties of the gradient, we know this direction is exactly opposite to the gradient vector, $-\nabla f$.

This simple idea is the foundation of the **Gradient Descent** algorithm, which is central to training modern machine learning models. The "altitude" of the hiker is the model's error (or loss function), and the "position" is the set of model parameters. The algorithm works as follows:

1. Start at a random point (random set of parameters).
2. Calculate the gradient of the loss function at that point.
3. Take a small step in the direction of the negative gradient, $-\nabla f$.
4. Repeat until the gradient is (close to) zero, indicating a local minimum has been reached.

This iterative process of following the negative gradient "downhill" allows us to find the optimal parameters that minimize a model's error, even when the function has millions of variables.

Example 2.4

A simple loss function is given by $L(w_1, w_2) = w_1^2 + 2w_2^2$. If we are at the point $(w_1, w_2) = (2, 1)$, in which direction should we move to decrease the loss fastest?

Solution: We need to find the direction of the negative gradient, $-\nabla L$. First, we compute the gradient:

$$\nabla L = \left\langle \frac{\partial L}{\partial w_1}, \frac{\partial L}{\partial w_2} \right\rangle = \langle 2w_1, 4w_2 \rangle$$

Next, we evaluate the gradient at the point $(2, 1)$:

$$\nabla L(2, 1) = \langle 2(2), 4(1) \rangle = \langle 4, 4 \rangle$$

The direction of steepest ascent is $\langle 4, 4 \rangle$. Therefore, the direction of steepest descent is the negative of this vector:

$$-\nabla L(2, 1) = \langle -4, -4 \rangle$$

To minimize the loss, we should adjust our parameters w_1 and w_2 in the direction $\langle -4, -4 \rangle$. ◀

2.4 Application: Training a Model with Gradient Descent

We have seen that the gradient, ∇f, points in the direction of steepest ascent, and its negative, $-\nabla f$, points in the direction of steepest descent. This principle is not merely a geometric curiosity; it is the engine that drives the training of most modern machine learning models. We will now explore a complete, practical example: finding the optimal line to fit a dataset using **Linear Regression** and the **Gradient Descent** algorithm.

Example 2.5 Optimising a Simple Linear Regression Model

Solution: Imagine we have a dataset consisting of n points (x_i, y_i). For instance, x_i could be the size of a house and y_i its price. Our goal is to find the straight line that best fits this data. This process is known as linear regression.

Step 1: Define the Model

A straight line is defined by its intercept (β_0) and its slope (β_1). For any given input x_i, our model predicts a value, which we'll call $\hat{y}_i$, according to the equation:

$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

Our task is to find the optimal values for the parameters β_0 and β_1.

Step 2: Define the Loss Function

How do we know if a line is a "good fit"? We measure its error. For each data point, the error is the difference between the actual value (y_i) and the predicted value ($\hat{y}_i$). To ensure errors don't cancel each other out and to

penalize larger errors more heavily, we square these differences. We then average this squared error over all n data points. This gives us the **Mean Squared Error (MSE)** loss function, L:

$$L(\beta_0, \beta_1) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

This is the central equation. Notice that the loss L is not a function of x or y (which are fixed data points), but a function of our model's parameters, β_0 and β_1. This loss function can be visualized as a convex, bowl-shaped surface where the vertical axis is the loss and the horizontal axes are the parameters β_0 and β_1, just as we saw with functions like $f(x, y)$ earlier. Our goal is to find the bottom of this bowl.

Step 3: Compute the Gradient of the Loss Function

To find the bottom of the loss surface using gradient descent, we need to know the direction of steepest descent at any point. This requires computing the gradient of L with respect to its variables, β_0 and β_1. The gradient is the vector of partial derivatives:

$$\nabla L = \left\langle \frac{\partial L}{\partial \beta_0}, \frac{\partial L}{\partial \beta_1} \right\rangle$$

We find each partial derivative using the rules of differentiation, particularly the chain rule.

For $\frac{\partial L}{\partial \beta_0}$, we treat β_1 as a constant and differentiate:

$$\frac{\partial L}{\partial \beta_0} = \frac{1}{n}\sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i) \cdot (-1) = -\frac{2}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)$$

For $\frac{\partial L}{\partial \beta_1}$, we treat β_0 as a constant and differentiate:

$$\frac{\partial L}{\partial \beta_1} = \frac{1}{n}\sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i) \cdot (-x_i) = -\frac{2}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i$$

These two expressions give us the components of the gradient vector ∇L for any given β_0 and β_1.

Step 4: The Gradient Descent Algorithm

Now we have all the pieces. The Gradient Descent algorithm is an iterative process:

1. Start with an initial guess for β_0 and β_1 (e.g., random values).
2. Calculate the gradient ∇L using the current values of β_0 and β_1.
3. Update the parameters by taking a small step in the direction of the *negative* gradient.

This update step is the core of the algorithm. We introduce a small positive constant η (eta), called the **learning rate**, which controls the size of our steps. The update rules are:

$$\beta_0 \leftarrow \beta_0 - \eta\frac{\partial L}{\partial \beta_0}$$

$$\beta_1 \leftarrow \beta_1 - \eta\frac{\partial L}{\partial \beta_1}$$

We repeat this process many times. With each step, our parameters β_0 and β_1 move closer to the values that minimize the loss function L, effectively "walking downhill" on the loss surface until they settle at the bottom.

Using more compact vector notation, where $\boldsymbol{\beta}$ is the vector $\langle \beta_0, \beta_1 \rangle$, the entire update rule can be written as:

$$\boldsymbol{\beta} \leftarrow \boldsymbol{\beta} - \eta \nabla L$$

This single expression elegantly captures the process of optimizing a machine learning model. It is a direct and powerful application of the gradient, demonstrating how a fundamental concept from calculus is used to solve complex, real-world problems. ◀

This chapter has extended the concept of the derivative into multiple dimensions. We started by exploring functions of several variables, partial derivatives, and finally the gradient vector. The gradient, as the "full" derivative for multivariable functions, not only describes the rate of change in all directions but also provides the theoretical foundation for powerful optimisation algorithms that drive modern technology. We then applied the gradient to the problem of finding the optimal line to fit a dataset using linear regression and the gradient descent algorithm. This concludes the chapter and effectively concludes this book about Mathematics for Software Engineering.

Bibliography

Cormen, T., Leiserson, C., Rivest, R., & Stein, C. (2022). *Introduction to algorithms* (4th). MIT Press.

Lay, D. (2003). *Linear algebra and its applications*. Pearson Education.

Montgomery, D. (2013). *Applied statistics and probability for engineers, 6th edition*. John Wiley; Sons, Incorporated.

Pishro-Nik, H. (2014). *Introduction to probability, statistics, and random processes*. Kappa Research, LLC.

Rosen, K. H. (2012). *Discrete mathematics and its applications* (7th). McGraw-Hill Education.

Ross, S. M. (2020). *Introduction to probability and statistics for engineers and scientists* (6th). Academic Press.

Thomas, C. *Introduction to differential calculus* [Accessed online]. Mathematics Learning Centre, University of Sydney. Accessed online. Sydney, NSW, 1997.

Appendix A: Important Concepts

This appendix is a collection of important mathematical concepts that are frequently used in software engineering. The content of this appendix is based on the concepts in this book.

Proposition A.1 (Order of Operations)

To evaluate mathematical expressions, operations are performed in the following order:

1. **Brackets (Parentheses):** First, perform all operations inside brackets or parentheses.
2. **Exponents and Radicals:** Next, evaluate exponents (powers) and radicals (roots).
3. **Multiplication and Division:** Then, perform multiplication and division from left to right.
4. **Addition and Subtraction:** Finally, execute addition and subtraction from left to right.

♠

Proposition A.2 (Rules for Calculations with Fractions)

For $a, b, c, m \in \mathbb{R}$, with $a, b, c, m \neq 0$ where required, the following identities hold:

(1) $\frac{a}{b} \times m = \frac{am}{b}$

(2) $\frac{a}{b} \div m = \frac{a}{bm}$

(3) $m \div \frac{a}{b} = \frac{mb}{a}$

(4) $\frac{a}{b} \times \frac{c}{a} = \frac{c}{b}$

(5) $\frac{a}{b} \div \frac{c}{a} = \frac{a^2}{bc}$

(6) $\frac{a}{b} = \frac{ac}{bc}$

(7) $\frac{a}{b} + \frac{c}{a} = \frac{a^2 + bc}{ab}$

♠

Proposition A.3 (Properties of Integer Exponents)

Let $n, m \in \mathbb{Z}$. Then the following hold (with $x, y \in \mathbb{R}$ and nonzero where stated):

(1) $x^n \cdot x^m = x^{n+m}$,

(2) $\frac{x^n}{x^m} = x^{n-m}$ with $x \neq 0$,

(3) $x^n \cdot y^n = (xy)^n$,

(4) $\frac{x^n}{y^n} = \left(\frac{x}{y}\right)^n$ with $y \neq 0$,

(5) $(x^n)^m = x^{nm}$,

(6) $x^1 = x$.

♠

Proposition A.4 (More Properties of Integer Exponents)

Let $n, m \in \mathbb{Z}$. Then the following hold (with $x, y \in \mathbb{R}$ and nonzero where stated):

(7) $x^0 = 1$ $\qquad x \neq 0$

(8) $\dfrac{1}{x^m} = x^{-m}$ $\qquad x \neq 0$

♠

Rules for rearranging formulae

The following operations can be performed on both sides of the formula:

- Add the same quantity to both sides
- Subtract the same quantity from both sides
- Multiply both sides by the same quantity - remember to multiply all terms
- Divide both sides by the same quantity - remember to divide all terms
- Apply a function to both sides, such as squaring or finding the reciprocal

Definition A.1 (Injective and Surjective Functions)

A function $f : A \to B$ is called **one-to-one** (or **injective**) if different elements in A map to different elements in B. A function $f : A \to B$ is called **onto** (or **surjective**) if every element in B is the image of at least one element in A.

♣

Definition A.2 (Inverse Functions)

Let f be a one-to-one correspondence from the set A to the set B. The inverse function of f is the function that assigns to an element b belonging to B the unique element a in A such that $f(a) = b$. The inverse function of f is denoted by f^{-1}. Hence, $f^{-1}(b) = a$ when $f(a) = b$.

♣

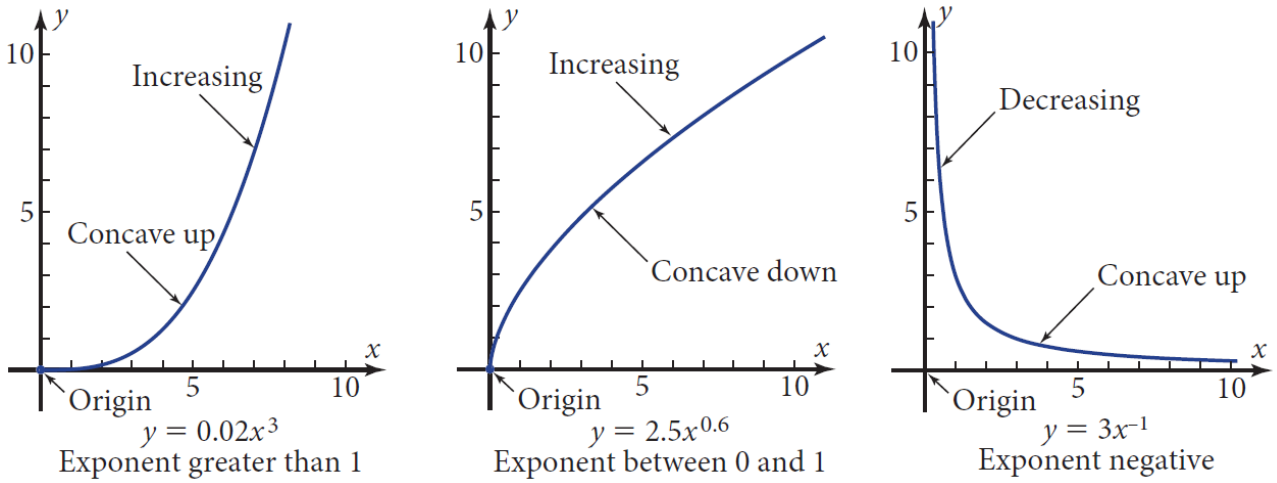


Figure A.1: Power functions

Definition A.3 (Base-10 Logarithms)

$$\log x = y \iff 10^y = x$$

Verbally: $\log x$ is the exponent in the power of 10 that gives x

♣

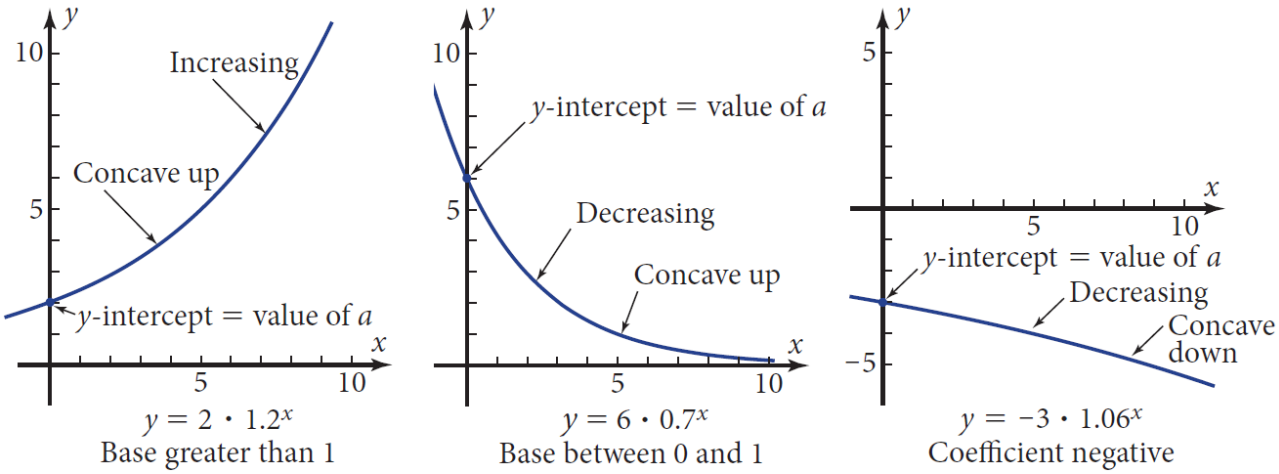


Figure A.2: Exponential functions

Properties of base-10 logarithms

- Log of a Product:

  $$\log xy = \log x + \log y$$

  Verbally: The log of a product equals the sum of the logs of the factors.

- Log of a Quotient:

  $$\log \frac{x}{y} = \log x - \log y$$

  Verbally: The log of a quotient equals the log of the numerator minus the log of the denominator.

- Log of a Power:

  $$\log x^y = y \log x$$

  Verbally: The log of a power equals the exponent times the log of the base.

Definition A.4 (Common Logarithm and Natural Logarithm)

Common: The symbol $\log x$ means $\log_{10} x$.

Natural: The symbol $\ln x$ means $\log_e x$, where e is a constant equal to $2.71828182845\ldots$

The Change-of-Base Property of Logarithms

$$\log_a x = \frac{\log_b x}{\log_b a} \quad \text{or} \quad \log_a x = \frac{1}{\log_b a}(\log_b x)$$

Properties of Logarithms

The Logarithm of a Power:

$$\log_b x^y = y \log_b x$$

The Logarithm of a Product:

$$\log_b(xy) = \log_b x + \log_b y$$

The Logarithm of a Quotient:

$$\log_b \frac{x}{y} = \log_b x - \log_b y$$

Numeral system	Symbols	Base	Additional information
Decimal	0-9	10	-
Binary	0, 1	2	-
Hexadecimal	0-9, A-F	16	$\mathrm{A} \equiv 10, \mathrm{B} \equiv 11, \mathrm{C} \equiv 12, \mathrm{D} \equiv 13, \mathrm{E} \equiv 14, \mathrm{F} \equiv 15$
Octal	0-7	8	-

Table A.1: Summary of Common Numeral Systems

Decimal number		In powers of 2	Power of 2				Binary number
			3	2	1	0	
8	=	2^3	1	0	0	0	1000
7	=	$2^2 + 2^1 + 2^0$	0	1	1	1	111
6	=	$2^2 + 2^1$	0	1	1	0	110
5	=	$2^2 + 2^0$	0	1	0	1	101
4	=	2^2	0	1	0	0	100
3	=	$2^1 + 2^0$	0	0	1	1	11
2	=	2^1	0	0	1	0	10
1	=	2^0	0	0	0	1	1

Table A.2: Decimal Numbers in Binary Representation

Proposition A.5 (Binary Addition Rules)

$0 + 0 = 0, \quad 0 + 1 = 1, \quad 1 + 0 = 1, \quad 1 + 1 = 10$

Proposition A.6 (Binary Multiplication Rules)

$0 \times 0 = 0$

$0 \times 1 = 0$

$1 \times 0 = 0$

$1 \times 1 = 1$

$1 \times 10_2 = 10_2$ (multiplying by base 10_2 adds a 0 to the end)

Proposition A.7 (XOR Operation)

XOR produces a 1 if the two bits being compared are different and a 0 if they are the same:

$$0 \oplus 0 = 0, \quad 0 \oplus 1 = 1, \quad 1 \oplus 0 = 1, \quad 1 \oplus 1 = 0$$

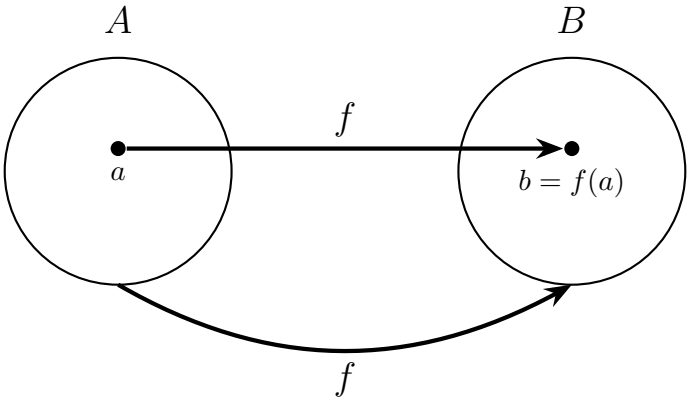


Figure A.3: A function f mapping an element a from set A to an element $b = f(a)$ in set B.

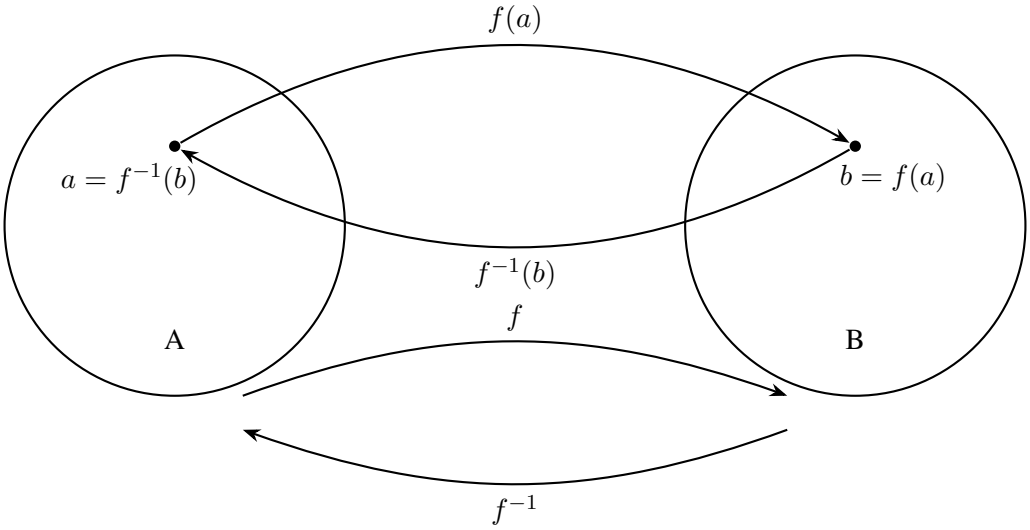


Figure A.4: The function f^{-1} is the inverse of function f.

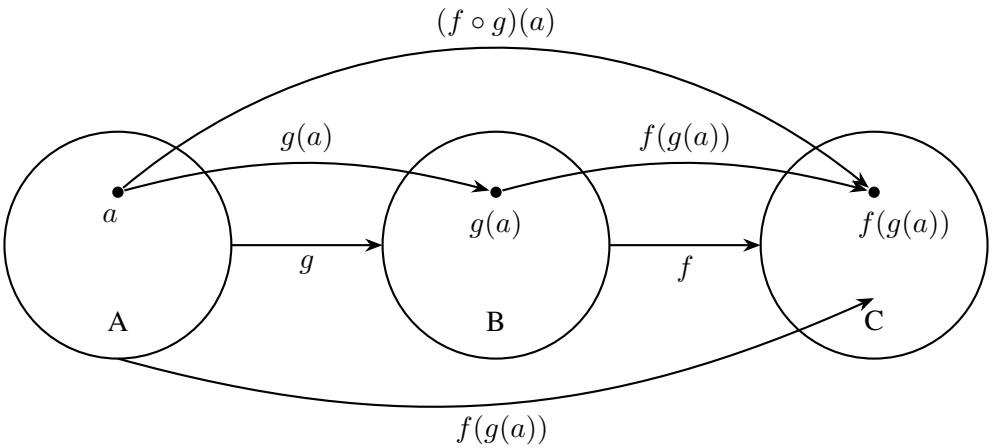


Figure A.5: The composition of functions f and g, denoted $f \circ g$, is the function that results from applying g and then f.